Continuous Expressive Speaking Styles Synthesis based on CVSM and MR-HMM

Jaime Lorenzo-Trueba [1,3]
Speech Technology Group[1]
Universidad Politecnica de Madrid
Spain
jaime@nii.ac.jp

Roberto Barra-Chicote [1]
Speech Technology Group[1]
Universidad Politecnica de Madrid
Spain

Ascension Gallardo-Antolin [2]
Signal Theory and Communications[2]
Universidad Carlos III de Madrid
Spain

Junichi Yamagishi [3]
National Institute of Informatics[3]
Tokyo
Japan

Juan M. Montero[1]
Speech Technology Group[1]
Universidad Politecnica de Madrid
Spain

Abstract

This paper introduces a continuous system capable of automatically producing the most adequate speaking style to synthesize a desired target text. This is done thanks to a joint modeling of the acoustic and lexical parameters of the speaker models by adapting the CVSM projection of the training texts using MR-HMM techniques. As such, we consider that as long as sufficient variety in the training data is available, we should be able to model a continuous lexical space into a continuous acoustic space. The proposed continuous automatic text to speech system was evaluated by means of a perceptual evaluation in order to compare them with traditional approaches to the task. The system proved to be capable of conveying the correct expressiveness (average adequacy of 3.6) with an expressive strength comparable to oracle traditional expressive speech synthesis (average of 3.6) although with a drop in speech quality mainly due to the semi-continuous nature of the data (average quality of 2.9). This means that the proposed system is capable of improving traditional neutral systems without requiring any additional user interaction.

1 Introduction

It is clear that in recent times there has been an increase in the penetration rates of speech technologies and applications based in speech recognition or speech synthesis are more and more common. Speech synthesis systems in particular have improved greatly in terms of speech intelligibility, speech quality or naturalness in neutral read speech situations regardless of the technology (Barra-Chicote, 2011). The problem appears when one needs to develop applications such as dialogue systems or robotic interfaces, for which a more expressive way of speaking is more appropriate.

One of the main problems regarding expressive speech synthesis is the vast amount of possibilities that have to be taken into account. Human expressiveness is not a discrete space but a continuous one, and speaking styles vary greatly from person to person and even from time period to time period. As such, obtaining enough data to cover all the requirements can become a very difficult task and scalability a significant problem. This is why statistical parametric speech synthesis is better fitted to the task. While unit-selection based systems have been proven to be more than capable of providing good quality

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 369–376, Osaka, Japan, December 11-17 2016.

expressive speech (Adell et al., 2012; Andersson et al., 2010), the adaptability of HMM-based systems allows us to better face the scalability problem.

The objective of the present paper is to introduce a complete TTS system capable of predicting the most adequate speaking style in order to automatically adapt the produced voice to the target text. For that objective we propose a continuous system based both in Continuous Vector Space Modeling (CVSM) (Tonta and Darvish, 2010; Klein et al., 2011) and Multi-regression HMM (MR-HMM) (Fujinaga et al., 2001), CVSM to model the expressive training texts as a continuous space and MR-HMM to make use of that continuous space as auxiliary features for the HMM modeling. This results in a system capable of characterizing input texts as a vector, which at the same time ends up producing the adequate acoustic model associated to the input text. As such, as long as sufficient variety in the training data is available, we would be able to model a continuous lexical space into a continuous acoustic space.

The rest of the paper is organized as follows. In section 2 we describe corpus considered for training the system. Section 3 gives a theoretical background to the proposed system and then explains in detail the proposed method. Then, section 4 describes the perceptual evaluations environment, whose results are described in section 4.1. Finally in section 5 we present the conclusions to be drawn from this paper together with some ideas for future work.

2 Speech Corpus

For the present research we wanted to combine text processing techniques and speech synthesis techniques, so the considered corpus had to not only cover a number of speaking styles with a reasonable amount of text data, but also consist of speech from a single speaker, so that the synthesis models could provide high quality synthetic speech. For that reason we utilized our self-designed and recorded database: Spanish Speaking Styles.

2.1 Spanish Speaking Styles

Spanish Speaking Styles (SSS) is a speaking styles corpus recorded for a single male professional speaker in 4 different speaking styles, with about 1 hour of speech per speaking style. The 4 speaking styles (news broadcasting, interviews, live sports broadcasts and political speech) cover a large spectrum of the expressive map (as can be seen in the F0-rhythm map that we can see in figure 1) while being recognizable.

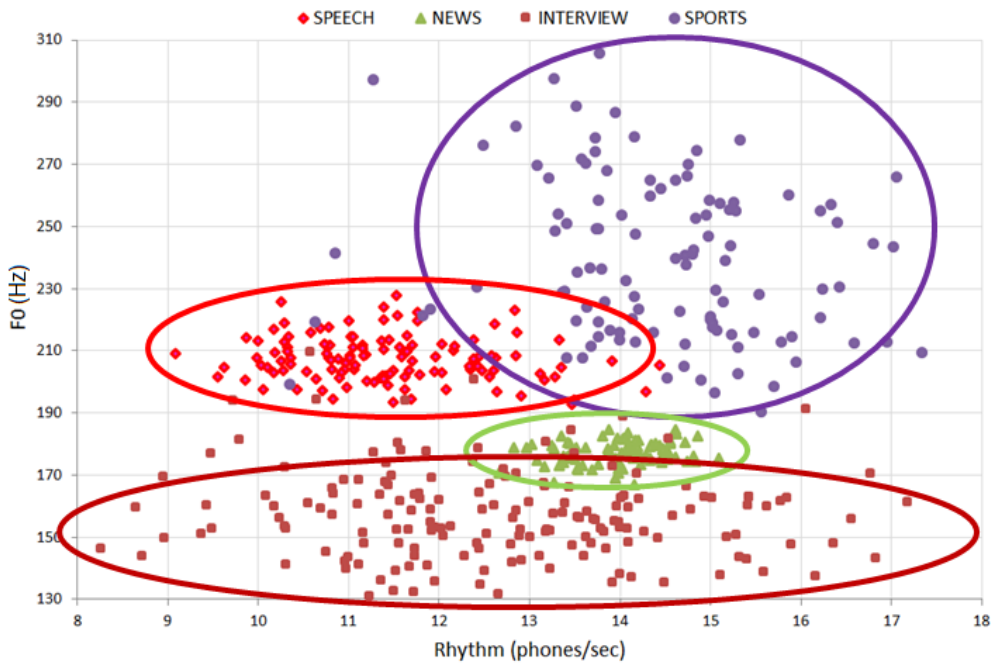

Figure 1: F0 vs. rhythm map of the 4 recorded speaking styles in SSS.

Two of the speaking styles (news and political speech) are scripted and the other two (live sports broadcasts and interviews) unscripted. A summary of the database can be seen in table 1.

Speaking Style	Train set	Test set
NEWS	102 utts, 1h3min	21 utts, 11min
INTERVIEW	332 utts, 45min	21 utts, 7min
SPORT	200 utts, 56min	21 utts, 7min
POLITICAL SPEECH	116 utts, 56min	21 utts, 11min
TOTAL	750 utts, 3h40min	84 utts, 36min

Table 1: Detailed description of the SSS database.

3 Expressive Speech Synthesis based on CVSM and MR-HMM

The ideal expressive TTS system should be able to handle any kind and any number of expressiveness without requiring costly labeling or manipulations every time we want to add any new one. With this purpose in mind the concept of a continuous system (i.e. a system in which expressiveness is treated as a complete space instead of as discrete entities) fits perfectly. That is, in real life expressiveness present overlap between them because there are not clear frontiers, so being able to take into account those overlaps in the shape of a dynamic synthesizer would be ideal.

3.1 Continuous Vector Space Modeling

Traditional information retrieval techniques such as TF-IDF (Fautsch and Savoy, 2010) are commonly based on the assumption that all the terms of the vocabulary lists of the documents do not have any relationship to each other, which is ultimately false as language has semantic relationships that we should be able to model (Tonta and Darvish, 2010; Klein et al., 2011). With that consideration in mind Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer et al., 2013), nowadays referred to as Continuous Vector Space Modeling (Krishnamurthy and Mitchell, 2013; Andreas and Ghahramani, 2013) was born.

CVSM aims to exploit the relationships between terms t_i and documents d_j by transforming them into an alternate "semantic" vector space, where both terms and documents are described by vectors of similar dimensionality and directions, enabling direct comparisons (Olmos et al., 2013; Cosma and Joy, 2012).

Typically this step is done by means of Singular Value Decomposition (SVD) (Golub and Reinsch, 1970; Henry et al., 2010), through which the latent semantic structure of the WTDM is shown.

3.2 Multi-Regression HMM

Multi-regression HMM is a particular kind of adaptation in which the model parameters are adapted depending on auxiliary features instead of the acoustic features themselves. Initially this was developed to exploit the correlation between F0 and the spectral features (Fujinaga et al., 2001), which significantly improved isolated word recognition rates in a speech recognition task. Numerically speaking, the multi-regression formula for M auxiliary features is defined as follows:

$$\mu = r_0 + r_1\epsilon_1 + \ldots + r_1\epsilon_M \tag{1}$$

Where $r_{0\ldots M}$ are the regression coefficients and $r_{\epsilon\ldots M}$ the M auxiliary features. This particularity can be exploited in speech synthesis to model additional information in the speaker models such as speaking styles or emotions (Nose and Kobayashi, 2012; Ling et al., 2013), and it has been used for applications as varied as generating walking motion models (Niwase et al., 2005).

3.3 Proposed System

Figure 2 shows the proposed flowchart for the continuous, MRHMM-based system. There we can see many fundamental differences with the discrete and semi-continuous approach, both at training time and at synthesis time.

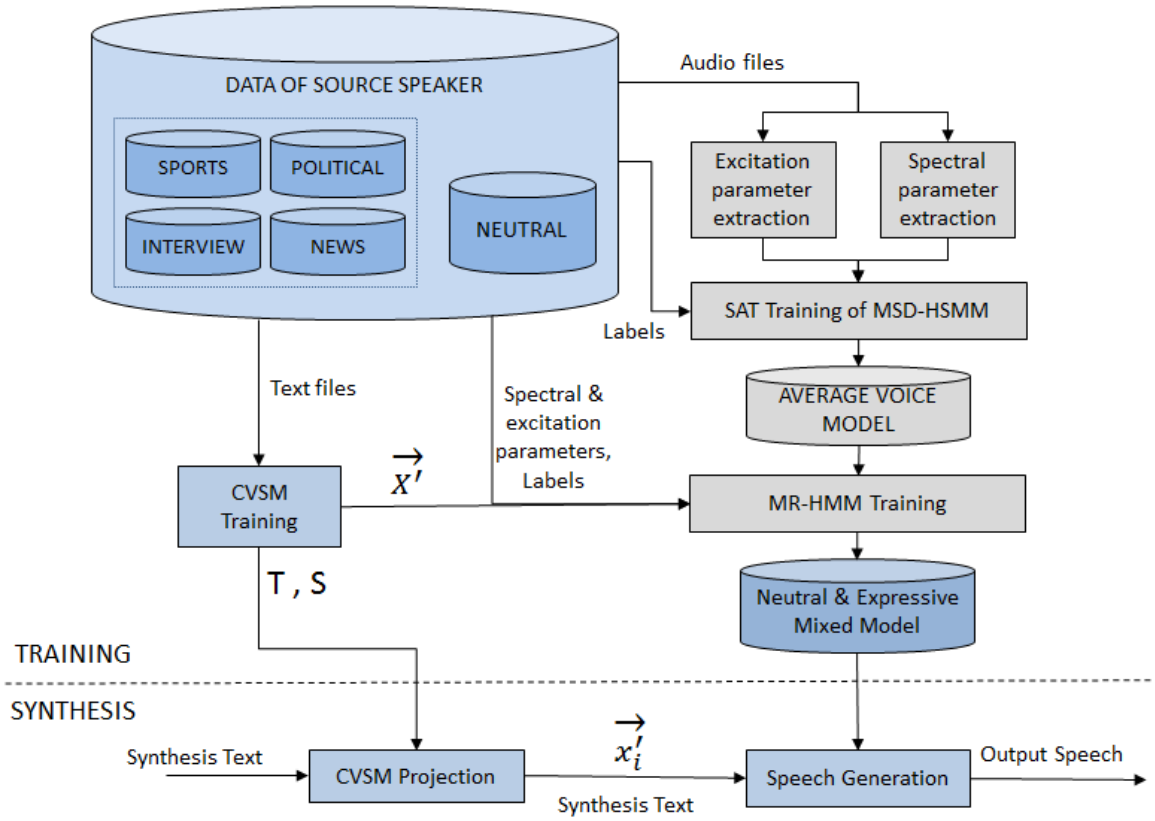

Figure 2: Schematic of the proposed continuous approach to the expressive TTS system.

Most notably, there is no traditional adaptation process involved in the system, but a MR-HMM adaptation that takes the CVSM projections of each training text file as the control vector $\vec{x}'$ for a training process that outputs a new model that combines both neutral and expressive information. Also we can see how there is no need for a centroid estimation, as there is no genre prediction carried out at synthesis time. Instead, only the CSVM matrices are kept as the output so that at synthesis time the synthesis text can be projected to obtain $\vec{x}'_i$, which will be used directly at the speech generation process. This process is then capable of producing an expressive speech output without relying in any genre prediction system, only in the CVSM projections of the synthesis texts.

4 Perceptual Evaluation

For the proposed perceptual evaluation we considered two approaches to the continuous modeling: a first one that directly utilized the CVSM-projected vectors $\vec{x}'$ and a second one that normalized each component by the maximum of their respective component $\vec{\bar{x}}' = \{v_1/v_{MAX1}, \ldots, v_4/v_{MAX4}\}$ where $v_{MAXi} = max\,(v_i)\,\forall v_i$ in the training data. The second approach was considered in case reducing the dispersion of the control vector values helped the MR-HMM re-estimation process. The same normalization was applied to the synthesis control vectors. In total 8 systems were evaluated (2 versions with 4 speaking styles each), which following the Latin Square approach (Gao, 2005), meant that we needed 8 different utterances to be synthesized for all the systems to be presented to the listeners in a random order without repetitions.

The test itself was carried out by means of a web interface, where the evaluator was presented with a button to play the audio sample and the transcription of the uttered texts. The samples could be played as many times as desired. Then, the listener was asked to rate the utterances in the traditional 5 point MOS evaluation in terms of adequacy of the utterance to the text (from not adequate to very adequate), speech quality (from very bad to very good) and perceived expressive strength when comparing to a hypothetical neutral version (very low to very high).

Regarding the speaker models, the whole train section of SSS database and the neutral speech of the same speaker present in SEV database (Barra-Chicote et al., 2008) were modeled into an average voice model (AVM) by applying Speaker Adaptive Training (SAT) (Anastasakos et al., 1997) with three feature streams with their Δ and Δ^2 coefficients: logarithm of the fundamental frequency (1 coefficient), mel-cepstral analysis coefficients (MCEP, 60 coefficients) and aperiodicity bands (25 coefficients). The models were adapted by means of the CSMAPLR algorithm (Yamagishi et al., 2009).

In terms of the statistical significance of the results, we applied the Wilkoxon Signed-Rank Test for a 95% confidence ratio in order to obtain the error margins. 16 subjects took part in each evaluation to guarantee double coverage of the Latin square matrix.

4.1 Evaluation Results

Figures 4 to 3 show the results for both continuous system evaluations (C_Normalized for the normalized version and Continuous for the non-normalized one), together with neutral speech (N), traditional emotional system (S) and natural voice (NAT). The results for the non-continuous systems have been extracted from other works to serve as a reference. It is important to emphasize that this evaluation considered only utterances that could be synthesized, which represented a 65% for the normalized version and 60% for the non-normalized one, details for each system-style interaction can be seen in table 2.

Synth. Rate	C_Normalized	Continuous
Interview	48%	43%
News	81%	67%
Speech	81%	90%
Sports	52%	38%
Average	65%	60%

Table 2: Percentage of synthesizable test sentences for each system-style pair. C_Normalized represents the normalized implementation of the continuous system and Continuous the non-normalized one.

Speech quality (figure 3) does show some bad results, an average quality of 2.93 for the basic system and 2.73 for the normalized system, which is significantly worse than all other systems. This is supposed to be mainly because the continuous system introduces a large amount of artifacts. This effect was somewhat expected due to the inherent semi-continuous nature of the SSS database: the professional speaker was asked to interpret the four speaking styles showing as little variation as possible so that they were clear representatives of their paralinguistic characteristics, so modeling them in a continuous fashion is not possible without additional data. More varied data or a more naturally continuous task is expected to fare better for the continuous system.

On the other hand, in the adequacy results (figure 4) we can see how the system provides significant increases when compared to traditional neutral systems. In the case of the normalized system, the average adequacy is 3.47 and 3.60 for the non-normalized version, not significantly worse than the 3.78 of the traditional expressive synthesis. Interviews, due to its extremely conversational nature was not modeled adequately.

Finally perceived expressive intensity results (figure 5) show a similar image to adequacy. Significantly better than using a neutral system, both proposed continuous systems show a 3.51 average perceived expressive intensity, which is once again comparable to the 3.65 of the non-predictive speech synthesis. Even so the systems are far from the 4.07 of natural speech and are much better than the 2.51 of neutral speech.

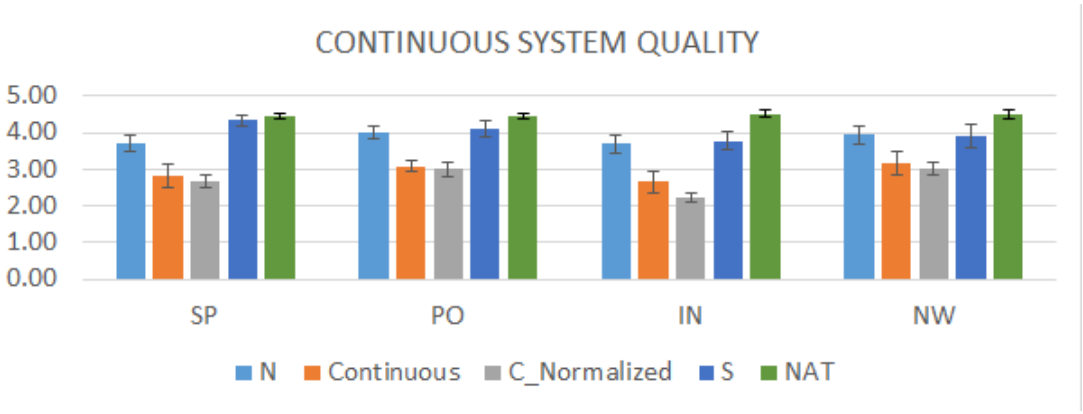

Figure 3: Results in MOS scale of the quality evaluation for the continuous system.

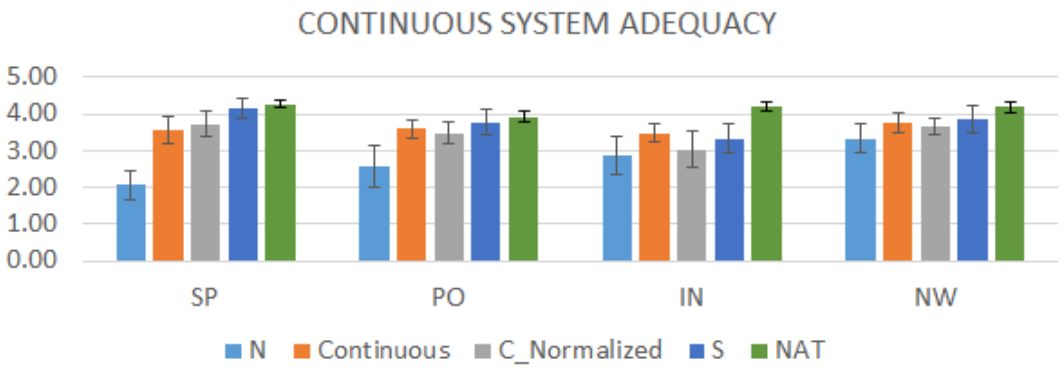

Figure 4: Results in MOS scale of the adequacy evaluation for the continuous system.

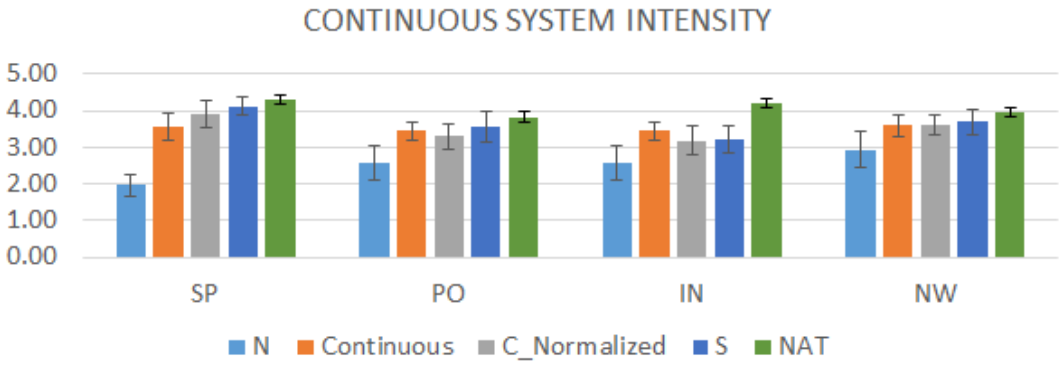

Figure 5: Results in MOS scale of the expressive intensity evaluation for the continuous system.

5 Conclusions and Future Work

We have introduced a complete TTS system is capable of synthesizing the most adequate expressiveness to the target text without relying in genre prediction techniques, just by making use of the CVSM projection of the input text as a control factor of the speaker model, which vastly increases the versatility of the system as we remove the need for labeling the training data genre. This system proved to be capable of conveying the correct expressiveness (average adequacy of 3.6) with an expressive strength comparable to oracle traditional expressive speech synthesis (average of 3.6) although at a significant drop in speech quality mainly due to the semi-continuous nature of the data (average quality of 2.9).

All in all we have introduced a different approach to expressive speech synthesis where the system automatically adjusts the produced speaking style according to the text to be synthesized without requiring any output from the user besides providing adequate training data. The system has shown that it is capable of significantly improving the traditional neutral speech synthesis systems in the task, and also of providing similar adequacy and perceived expressive strength rates than those of natural voice.

For future work we want to consider a broader array of expressiveness, in order to find a problem where the continuous modeling fits naturally: a more complete speaking styles collection, or even considering sub-spaces of speaking styles, including more conversational speaking styles in an attempt to solve the problems that arose with interviews. Another field that we want to work on is on bringing our systems into the DNNs and RNNs world. This task will prove challenging as there is still not many researches underway on DNN-based expressive speech synthesis. Finally, carrying out evaluations in real life systems such as car navigation systems or robotic assistants in scenarios not as constrained as the ones we evaluated would provide much needed information on how research systems fare in the real world, which would undoubtedly give hints on where more to focus our efforts.

Acknowledgements

The work leading to these results has received funding from the European Union under grant agreement 287678. It has been supported by NAVEGABLE (DPI2014-53525-C3-2-R), ASLP-MULN (TIN2014-54288-C4-1-R) projects and by Spanish Government grant TEC2014-53390-P. Jaime Lorenzo has been funded by Universidad Politécnica de Madrid under grant SBUPM-QTKTZHB. The authors want to specially thank the members of the Speech Technology Group, NAVEGASE and Simple4All, specially Simon and Oliver from CSTR, for the continuous and fruitful discussion on these topics.

References

Jordi Adell, David Escudero, and Antonio Bonafonte. 2012. Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, 54(3):459–476.

Tasos Anastasakos, John McDonough, and John Makhoul. 1997. Speaker adaptive training: A maximum likelihood approach to speaker normalization. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1043–1046. IEEE.

Sebastian Andersson, Kallirroi Georgila, David Traum, Matthew Aylett, and Robert AJ Clark. 2010. Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. *Speech Prosody*.

Jacob Andreas and Zoubin Ghahramani. 2013. A generative model of vector space semantics. *ACL 2013*, page 91.

R. Barra-Chicote, J. M. Montero, J. Macias-Guarasa, S. Lufti, J. M. Lucas, F. Fernandez, L. F. D'haro, R. San-Segundo, J. Ferreiros, R. Cordoba, and J. M. Pardo. 2008. Spanish expressive voices: Corpus for emotion research in spanish. *Proc. of LREC*.

Roberto Barra-Chicote. 2011. *Contributions to the analysis, design and evaluation of strategies for corpus-based emotional speech synthesis*. Ph.D. thesis, ETSIT-UPM.

Georgina Cosma and Mike Joy. 2012. An approach to source-code plagiarism detection and investigation using latent semantic analysis. *Computers, IEEE Transactions on*, 61(3):379–394.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Claire Fautsch and Jacques Savoy. 2010. Adapting the tf idf vector-space model to domain specific information retrieval. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1708–1712. ACM.

Katsuhisa Fujinaga, Mitsuru Nakai, Hiroshi Shimodaira, and Shigeki Sagayama. 2001. Multiple-regression hidden markov model. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 513–516. IEEE.

Lei Gao, 2005. *Latin Squares in Experimental Design*. Michigan State University.

Gene H Golub and Christian Reinsch. 1970. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420.

ER Henry, J Hofrichter, et al. 2010. Singular value decomposition: application to analysis of experimental data. *Essential Numerical Computer Methods*, 210:81–138.

Richard Klein, Angelo Kyrilov, and Mayya Tokman. 2011. Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*, pages 158–162. ACM.

Jayant Krishnamurthy and Tom M Mitchell. 2013. Vector space semantic parsing: A framework for compositional vector space models. *ACL 2013*, page 1.

Thomas K Landauer, Danielle S McNamara, Simon Dennis, and Walter Kintsch. 2013. *Handbook of latent semantic analysis*. Psychology Press.

Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi. 2013. Articulatory control of hmm-based parametric speech synthesis using feature-space-switched multiple regression. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(1):207–219.

Naotake Niwase, Junichi Yamagishi, and Takao Kobayashi. 2005. Human walking motion synthesis with desired pace and stride length based on hsmm. *IEICE transactions on information and systems*, 88(11):2492–2499.

Takashi Nose and Takao Kobayashi. 2012. An intuitive style control technique in hmm-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. *Speech Communication*.

Ricardo Olmos, José A León, Guillermo Jorge-Botana, and Inmaculada Escudero. 2013. Using latent semantic analysis to grade brief summaries: A study exploring texts at different academic levels. *Literary and linguistic computing*, 28(3):388–403.

Yaşar Tonta and Hamid R Darvish. 2010. Diffusion of latent semantic analysis as a research tool: A social network analysis approach. *Journal of Informetrics*, 4(2):166–174.

J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai. 2009. Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(1):66–83.